STATISTICS FOR ECONOMISTS: A BEGINNING

John E. Floyd
University of Toronto

July 2, 2010

PREFACE

The pages that follow contain the material presented in my introductory quantitative methods in economics class at the University of Toronto. They are designed to be used along with any reasonable statistics textbook. The most recent textbook for the course was James T. McClave, P. George Benson and Terry Sincich, *Statistics for Business and Economics*, Eighth Edition, Prentice Hall, 2001. The material draws upon earlier editions of that book as well as upon John Neter, William Wasserman and G. A. Whitmore, *Applied Statistics*, Fourth Edition, Allyn and Bacon, 1993, which was used previously and is now out of print. It is also consistent with Gerald Keller and Brian Warrack, *Statistics for Management and Economics*, Fifth Edition, Duxbury, 2000, which is the textbook used recently on the St. George Campus of the University of Toronto. The problems at the ends of the chapters are questions from mid-term and final exams at both the St. George and Mississauga campuses of the University of Toronto. They were set by Gordon Anderson, Lee Bailey, Greg Jump, Victor Yu and others including myself.

This manuscript should be useful for economics and business students enrolled in basic courses in statistics and, as well, for people who have studied statistics some time ago and need a review of what they are supposed to have learned. Indeed, one could learn statistics from scratch using this material alone, although those trying to do so may find the presentation somewhat compact, requiring slow and careful reading and thought as one goes along.

I would like to thank the above mentioned colleagues and, in addition, Adonis Yatchew, for helpful discussions over the years, and John Maheu for helping me clarify a number of points. I would especially like to thank Gordon Anderson, who I have bothered so frequently with questions that he deserves the status of mentor.

After the original version of this manuscript was completed, I received some detailed comments on Chapter 8 from Peter Westfall of Texas Tech University, enabling me to correct a number of errors. Such comments are much appreciated.

J. E. Floyd
July 2, 2010

Chapter 5

Tests of Hypotheses

In the previous chapter we used sample statistics to make point and interval estimates of population parameters. Often, however, we already have some theory or hypothesis about what the population parameters are and we need to use our sample statistics to determine whether or not it is reasonable to conclude that the theory or hypothesis is correct. Statistical procedures used to do this are called *statistical tests.*

Consider, for example, the case of a firm that has developed a diagnostic product for use by physicians in private practice and has to decide whether or not to mount a promotional campaign for the product. Suppose that the firm knows that such a campaign would lead to higher profits only if the mean number of units ordered per physician is greater than 5. Office demonstrations are conducted with a random sample of physicians in the target market in order to decide whether or not to undertake the campaign. The campaign is very costly and the firm will incur substantial losses if it undertakes it only to find that the mean number of orders after the campaign is less than or equal to 5.

5.1 The Null and Alternative Hypotheses

We can think of two possibilities. The mean number of orders in the population of all physicians will exceed 5 or the mean will not exceed 5. Suppose the firm accepts the hypothesis that the mean number of orders in the population will be greater than 5 when it turns out to be less. A promotional campaign will be conducted at great loss. Had the guess that the mean number of orders will be greater than 5 been correct the firm would have earned a substantial profit. Alternatively, if the firm accepts the hypothesis that the mean number of orders in the population will be less than 5 when it turns out to be greater, some profit will be foregone. Had the guess that the mean number of orders in the population will be less than 5 been correct, however, huge losses from the promotional campaign will have been avoided. It turns out that the cost of guessing that the mean number of orders will be greater than 5, mounting the promotional campaign, and being wrong is much greater than the cost of guessing that the mean number of orders will be less than or equal to 5, not mounting the promotional campaign, and being wrong.

We call the more serious of the two possible errors a *Type I error* and the least serious error a *Type II error*. We call the hypothesis which *if wrongly rejected* would lead to the more serious (Type I) error the *null hypothesis* and denote it by the symbol H_0. The other hypothesis, which if wrongly rejected would lead to the less serious (Type II) error, we call the *alternative hypothesis* and denote it by the symbol H_1.

In the problem we have been discussing, the most serious error will occur if the mean number of orders in the population of physicians will be less than 5 and the firm erroneously concludes that it will be greater than 5. Hence, the null hypothesis is

$$H_0\colon\ \mu \leq 5$$

and the alternative hypothesis is

$$H_1\colon\ \mu > 5.$$

Acceptance of either hypothesis on the basis of sample evidence involves a risk, since the hypothesis chosen might be the incorrect one. We denote the probability of making a Type I error (incorrectly rejecting the null hypothesis) an α-risk and the probability of making a Type II error (incorrectly rejecting the alternative hypothesis) a β-risk. It turns out that if the sample size is predetermined (i.e., beyond the firm's control) the firm has to choose which risk to control. Control of the α-risk at a lower level will imply a greater degree of β-risk and vice versa. Since by construction Type I errors are the most damaging, the firm will obviously want to control the α-risk.

Of course, the situation could have been different. The market for the type of diagnostic product that the firm has developed may be such that the first firm providing it could achieve quite an advantage. An erroneous conclusion by the firm that the mean number of orders will be less than 5, and the resulting decision not to promote the product, could lead to the loss of substantial future market opportunities. On the other hand, if the cost of the promotion is small, an erroneous conclusion that the number of orders per physician in the population will equal or exceed 5 would perhaps lead to a minor loss. In this case we would define the null and alternative hypotheses as

$$H_0\colon\ \mu \geq 5$$

and

$$H_1\colon\ \mu < 5.$$

A Type I error will then result when the null hypothesis is incorrectly rejected—i.e., when we erroneously conclude that the mean order per physician in the population will be less than 5 when it turns out to be equal to or greater than 5. The probability of this happening will be the α-risk. A Type II error will result when the alternative hypothesis is incorrectly rejected—i.e., when the firm erroneously concludes that the mean order per physician will be greater than or equal to 5 when it turns out not to be. The probability of this happening will be the β-risk.

The hypotheses in the above problem were one-sided alternatives. The crucial question was whether the population parameter μ was above a particular value μ_0 $(= 5)$ or below it. We can also have two sided alternatives.

Suppose it is found that the mean duration of failed marriages was 8.1 years before the divorce law was changed and we want to determine whether the new legislation has affected the length of time unsuccessful marriages drag on. A sociologist has a random sample of divorce records accumulated since the law was changed upon which to make a decision. Erroneously concluding that the new legislation has changed people's behaviour when it has not is judged to be a more serious error than incorrectly concluding that behaviour has not changed as a result of the new law when it in fact has. Accordingly, the sociologist chooses the null hypothesis as

$$H_0\colon\ \mu = 0$$

and the alternative hypothesis as

$$H_1\colon\ \mu \neq 0.$$

A Type I error will arise if the sociologist concludes that behaviour has changed when it has not—i.e., incorrectly rejects the null hypothesis—and a Type II error will arise if she erroneously concludes that behaviour has not changed when it in fact has. The probability of a Type I error will again be the α-risk and the probability of a Type II error the β-risk.

5.2 Statistical Decision Rules

Take the case of the diagnostic product discussed above where H_0 is $\mu \leq 5$ and H_1 is $\mu > 5$. If upon conducting the office demonstrations the mean number of orders of physicians in the sample is less than 5, it would be reasonable to accept the null hypothesis that $\mu \leq 5$. If the sample mean is greater than 5, however, should we reject the null hypothesis? Clearly, the costs of a Type I error are greater than the costs of a Type II error, so we would not want to reject the null hypothesis if the sample mean is just a little bit above 5 because the sample mean could be greater than 5 entirely as a result of sampling error. On the other hand, if the sample mean is 20, it might seem reasonable to reject the null hypothesis. The question is: At what value of the sample mean should we reject the null hypothesis that $\mu \leq 5$. That value of the mean (or *test statistic*) at which we decide (ahead of time, before the sample is taken) to reject the null hypothesis is called the *action limit* or *critical value.* The choice of this critical value is called a *statistical decision rule.*

The general form of the statistical decision rule for one-sided and two-sided alternatives is given in Figure 5.1. Possible values of the sample mean are divided into two groups along the continuum of values the sample mean can take. The groups are separated by the critical value A in the case of one-sided tests shown in the top two panels, or by the critical values A_1 and A_2 in the case of a two-sided test shown in the bottom panel. The region between the critical value or values and μ_0, the level of μ at which the test is being conducted, is called the *acceptance region.* The region on the other side(s) of the critical value(s) from μ_0 is called the *critical region* or *rejection region.* If the sample mean falls in the rejection region, we reject the null hypothesis and accept the alternative hypothesis. If it falls in the acceptance region we accept the null hypothesis and reject the alternative hypothesis. Note that acceptance of the null hypothesis means only that we will act *as if* it were true—it does not mean that the null hypothesis is in fact true.

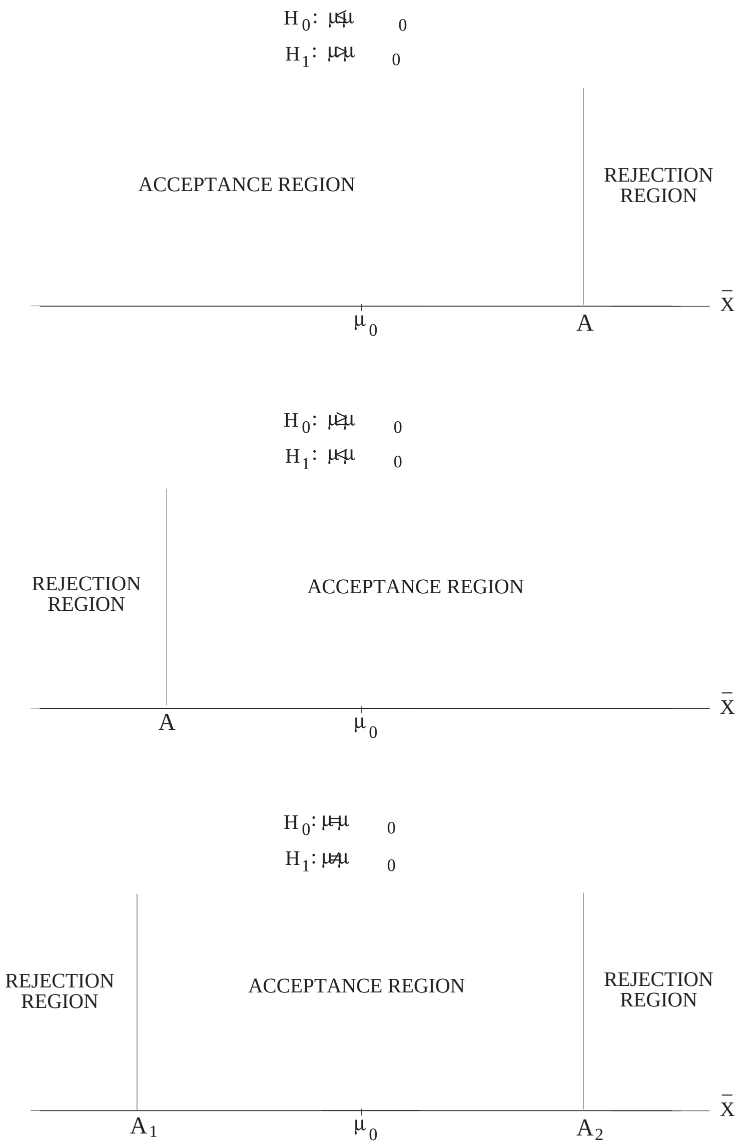

Figure 5.1: Ilustration of statistical decision rules for one-sided upper-tail (top), one-sided lower-tail (middle) and two-sided (bottom) alternatives concerning the population mean μ.

5.3 Application of Statistical Decision Rules

In order to actually perform the statistical test we must establish the degree of α-risk (risk of erroneously rejecting the null hypothesis) we are willing to bear. We must also make sure we are satisfied with the level of μ at which the α-risk is to be controlled—that is, with the level at which we set μ_0. In the example of the diagnostic product, we need not have set the level of μ at which the α-risk is to be controlled at 5. We could have been safer (in the case where the most costly error is to incorrectly conclude that the mean number of orders from the population of physicians is greater than 5 when it is in fact less than or equal to 5) to control the α-risk at $\mu_0 = 5.5$. At any given level of α-risk chosen this would have been a more stringent test than setting μ_0 at 5. We also have to establish the probability distribution of the standardised random variable $(\bar{X} - \mu)/s_{\bar{x}}$. If the sample size is large, the Central Limit Theorem tells us that it will be approximately normally distributed. If the sample size is small and the probability distribution of the population values X_i around μ is not too different from the normal distribution, $(\bar{X} - \mu)/s_{\bar{x}}$ will follow a t-distribution.

Suppose an airline takes a random sample of 100 days' reservation records which yields a mean number of no-shows on the daily flight to New York City of 1.5 and a value of s equal to 1.185. The resulting value of $s_{\bar{x}}$ is $1.185/\sqrt{100} = 1.185/10 = .1185$. The airline knows from extensive experience that the mean number of no-shows on other commuter flights is 1.32. The airline wants to test whether the mean number of no-shows on the 4 PM flight exceeds 1.32. We let H_0 be the null hypothesis that the mean number of no-shows is less than or equal to 1.320 and the alternative hypothesis H_1 be that the mean number of no shows exceeds 1.320. Notice that the hypothesis is about the number of no-shows in the whole population of reservations for the 4 PM flight to New York City. The airline wants to control the α-risk at .05 when $\mu = 1.320$. Since the sample is large

$$z = \frac{\bar{X} - \mu_0}{s_{\bar{x}}}$$

is approximately standard normal. The sample results in a value of z equal to

$$z^* = \frac{1.500 - 1.320}{.1185} = 1.519.$$

At an α-risk of .05 the critical value for z is 1.645 in a one-sided test. Thus, since z^* is less than the critical value we cannot reject the null hypothesis. We accept H_0 and reject H_1 since the standardised value of the sample mean does not fall in the critical region. The probability of observing a sample mean of 1.50 when the population mean is 1.320 is more than .05. This is an example of a one-sided upper-tail test because the critical region lies in the upper tail of the distribution. For an example of a one-sided lower-tail test consider a situation where a customs department asks travellers returning from abroad to declare the value of the goods they are bringing into the country.

The authorities want to test whether the mean reporting error is negative—that is, whether travellers cheat by underreporting. They set the null hypothesis as H_0: $\mu \geq 0$ and the alternative hypothesis as H_1: $\mu < 0$. A random sample of 300 travellers yields $\bar{X} = -\$35.41$ and $s = \$45.94$. This implies $s_{\bar{x}} = 45.94/17.32 = 2.652$. The α-risk is to be controlled at $\mu_0 = 0$. The sample size is again so large that the test statistic is distributed approximately as the standardised normal distribution. The sample yields a value equal to

$$z^* = \frac{-35.41 - 0}{2.652} = -13.35.$$

The authorities want to control the α-risk at .001 so the critical value for z is -3.090. Since z^* is well within the critical region we can reject the null hypothesis H_0 that the mean reporting error is non-negative and accept the alternative hypothesis that it is negative. In fact, the observed sample mean is 13.35 standard deviations below the hypothesized population mean of zero while the critical value is only 3.090 standard deviations below zero. Note that the α-risk is only approximately .001 because z is only approximately normally distributed.

Now let us take an example of a two-sided test. Suppose that a random sample of 11 children out of a large group attending a particular camp are given a standard intelligence test. It is known that children of that age have mean scores of 100 on this particular test. The camp organizers want to know whether or not the children attending the camp are on average equal in intelligence to those in the population as a whole. Note that the relevant population here from which the sample is drawn is the entire group of children attending the camp. The sample mean score was $\bar{X} = 110$ and s was equal to 8.8, resulting in a value for $s_{\bar{x}}$ of $8.8/3.62 = 2.65$. Since the concern is about possible differences in intelligence in either direction the appropriate test is a two-tailed test of the null hypothesis H_0: $\mu = \mu_0 = 100$ against the alternative hypothesis H_1: $\mu \neq \mu_0 = 100$. With a small sample size, under the assumption that the distribution of the population is not too far from normal,

$$\frac{\bar{X}-\mu}{s_{\bar{x}}}$$

will be distributed according to the t-distribution with 10 degrees of freedom. Suppose that the organizers of the camp want to control the α-risk at .05 at a value of $\mu_0 = 100$. Since the test is a two-tailed test the critical region has two parts, one at each end of the distribution, each containing probability weight $\alpha/2 = .025$ (the two together must have probability weight .05). This two-part region will contain those t-values greater than 2.228 and less than -2.228. The value of t that arises from the sample,

$$t^* = \frac{110-100}{2.65} = 3.77$$

clearly lies in the upper part of the critical region so that the null hypothesis that the intelligence level of the children in the camp is the same as that of those in the population as a whole must be rejected.

The decision rules for tests of μ can be shown in Figure 5.2. In the upper panel, which illustrates a one-sided upper-tail test, α is the probability that $\bar{X}$ will fall in the critical region if $\mu \leq \mu_0$. The area $1-\alpha$ is the probability that $\bar{X}$ will fall in the acceptance region. If $\bar{X}$ in fact falls in the rejection region, the probability will be less than α of observing that value, given the sample size, if μ is really less than or equal to μ_0. The center panel does the same thing for a one-sided lower-tail test. Here, $\bar{X}$ must fall below A for the null hypothesis to be rejected. The bottom panel presents an illustration of a two-sided test. The null hypothesis is rejected if $\bar{X}$ falls either below A_1 or above A_2. The probability of rejecting the null hypothesis if $\mu = \mu_0$ is equal to $\alpha/2 + \alpha/2 = \alpha$. We reject the null hypothesis if the probability of observing a sample mean as extreme as the one we obtain conditional upon $\mu = \mu_0$ is less than α.

5.4 P–Values

In the statistical test involving the average intelligence of children at the camp the value of z that resulted from the sample was 3.77 whereas the critical value was ±2.228. The probability of obtaining this sample from a population of children having mean intelligence of 100 is less than .05. An appropriate question is: What is the probability of observing a sample mean as extreme as the one observed if the mean intelligence of the population of children at the camp is 100? Or, to put it another way, what level of

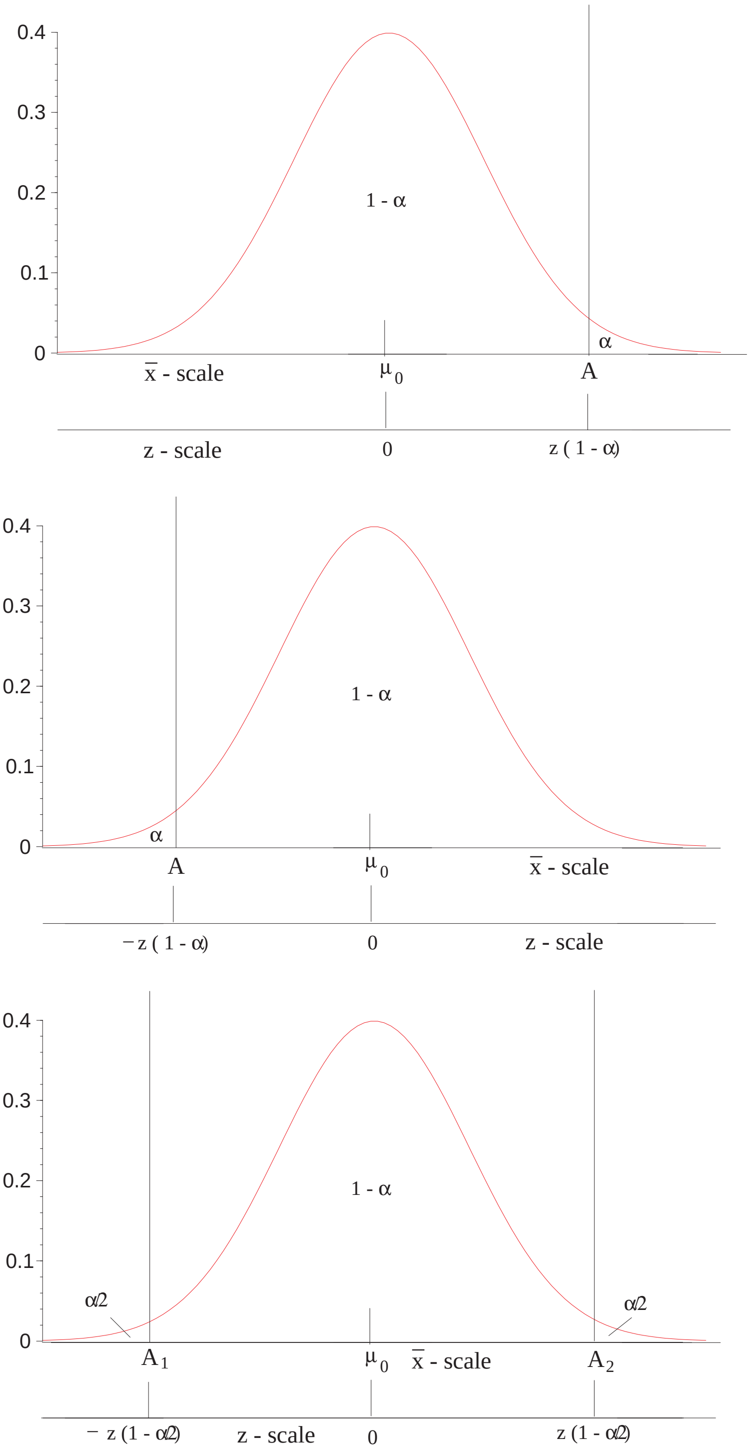

Figure 5.2: Ilustration of hypothesis tests—one-sided upper-tail (top), one-sided lower-tail (middle) and two-sided (bottom).

α-risk would have had to be selected for a borderline rejection of the null hypothesis? This probability is called the P–value. Formally, the P–value of a statistical test for μ is the probability that, if $\mu = \mu_0$, the standardised test statistic z might have been more extreme in the direction of the rejection region than was actually observed.

In the case of the children's intelligence, $\alpha/2$ would have had to be about .00275 for $t = 3.77$ to pass into the right rejection region of the t-distribution. Since the test is a two-sided one, the α-risk will be two times .00275 or .0055. The P–value is thus .0055 or somewhat more than half of one percent.

In the case of the customs department example, the value of z of roughly -13 is so far beyond the critical value of -2.28 that the α-risk required to get us to borderline reject the null hypothesis would be miniscule. Note that in this case there is only one critical region because the test is a one-tailed test, so we do not double the probability weight in that region to obtain the P–value.

The case of the no-shows on the commuter flight to New York City is more interesting because the value of z obtained from the sample is slightly less than the critical value of 1.645 when the α-risk is set at .05. The associated P–value equals

$$P(\bar{X} > 1.50|\mu = 1.32) = P(z > 1.519) = .0643.$$

There is a bit better than a 6 percent chance that we could have as many no-shows in a sample of 100 if the true mean number of no-shows on the 4 PM flight is 1.32, the mean number of no-shows on all flights.

In Figure 5.2 the P–Value would be the area to the right of our actual sample mean in the upper panel, the area to the left of our actual sample mean in the middle panel, and twice the smaller of the areas to the right or left of the actual sample mean in the lower panel.

5.5 Tests of Hypotheses about Population Proportions

When the population parameter of interest is a proportion p and the sample size is large enough to permit a normal approximation to the relevant binomial distribution, the above results go through with little modification apart from the calculation of the standard deviation of the sample proportion $\bar{p}$. It was shown in equation (4.5) of the previous chapter that the $\bar{p}$ has variance

$$Var\{\bar{p}\} = \frac{p\,(1-p)}{n},$$

and standard deviation

$$s_{\bar{p}} = \sqrt{\frac{p\,(1-p)}{n}}.$$

For example, consider a situation where the proportion of workers who are chronically ill in a particular region is known to be .11, and a random sample of 1000 workers in one of the many industries in that region yields a sample proportion of chronically ill equal to .153. We want to test whether the population of workers in that particular industry contains a higher proportion of chronically ill than the proportion of chronically ill in the entire region. Since the worst possible error would be to erroneously conclude that the proportion of chronically ill workers in the industry is bigger than the proportion in the region, we let the null hypothesis be H_0: $p \leq .11$ and the alternative hypothesis be H_1: $p > .11$. If the null hypothesis is true the standard deviation of $\bar{p}$ will equal $\sqrt{(.11)(1-.11)/1000} = .009894$. The value of the test statistic then becomes

$$z^* = \frac{\bar{p}-p}{s_{\bar{p}}} = \frac{.153-.110}{.009894} = 4.35.$$

If we are willing to assume an α-risk of .01 in this one-sided upper-tail test the critical value of z would be 2.326. Since the sample statistic exceeds the critical value we reject the null hypothesis that the proportion of chronically ill workers in the industry is the same as or less than the proportion of chronically ill workers in the entire region.

5.6 Power of Test

Our decision rules for tests of μ have been set up to control the α-risk of the test when $\mu = \mu_0$. But we should not be indifferent about the β-risk—i.e., the risk of rejecting the alternative hypothesis when it is true. Tests that have a high risk of failing to accept the alternative hypothesis when it is true are said to have *low power.* So we now pose the question: How big is the β-risk?

Let us consider this question with in the framework of a practical problem. Suppose that the country-wide mean salary of members of a professional association is known to be \$55.5 thousand. A survey of 100 members of one of the provincial branches of the association found a mean salary in that province of $\bar{X} = \$62.1$ thousand with $s = \$24.9$ thousand, yielding $s_{\bar{x}} = 24.9/10 = \$2.49$ thousand. We want to determine whether the mean salary of members in the province in question exceeds the known mean salary of members country-wide. Let us set the α-risk at .05, controlled at $\mu_0 = 55.5$. The critical value of z is 1.645, yielding a value for A of

$$A = \mu + z(1 - \alpha)s_{\bar{x}} = \mu + z(.95)(2.49) = 55.5 + (1.645)(2.49) = 59.5965.$$

The sample statistic is 62.1, well above the critical value. The standardised sample statistic is

$$z^* = \frac{62.1 - 55.5}{2.49} = \frac{6.6}{2.49} = 2.65$$

which is, of course, well above 1.645. The P–Value of the sample statistic is

$$P(\bar{X} \geq 62.1) = P(z^* \geq 2.65) = (1 - P(z^* < 2.65)) = 1 - .996 = .004.$$

While the α-risk is .05 controlled at $\mu_0 = 55.5$, the β-risk will depend on where μ actually is. Suppose that μ is actually an infinitesimal amount above 55. The null hypothesis is then false and the alternative hypothesis is true. Given our critical value A, however, there is almost a .05 probability that we will reject the null hypothesis and accept the alternative hypothesis. This means that the probability we will reject the alternative hypothesis when it is in fact true—the β-risk—is very close to .95.

Now suppose that μ is actually 57.1. The true distribution of $\bar{X}$ is then centered on $\mu = 57.1$ in the second panel from the top in Figure 5.3. About 16.1% of the distribution will now lie above the critical value A, so the probability that we will reject the null hypothesis is .16. This probability is called the *rejection probability* or the *power of test*. The probability that we will reject the alternative hypothesis is now 1 - .16 = .84. This probability—the probability of rejecting the alternative hypothesis when it is true—is the β-risk.

Suppose, instead, that μ is actually 59.6. As can be seen from the second panel from the bottom of Figure 5.3 this implies that the distribution of the test statistic is centered around the critical value A. The probability that we will reject the null hypothesis and accept the alternative hypothesis (i.e., the rejection probability or the power of test) is now .5. And the β-risk is also .5 (unity minus the rejection probability).

Finally, suppose that μ is actually 64.5. The distribution of the test statistic will now be centered around this value and, as can be seen from the bottom panel of Figure 5.3, .975 of that distribution now lies in the rejection region. The power of test is now .975 and the β-risk equals (1 - .975) = .025.

So the higher the actual value of μ the greater is the power of test and the lower is the β-risk. This can be seen from Figure 5.4. The curve in

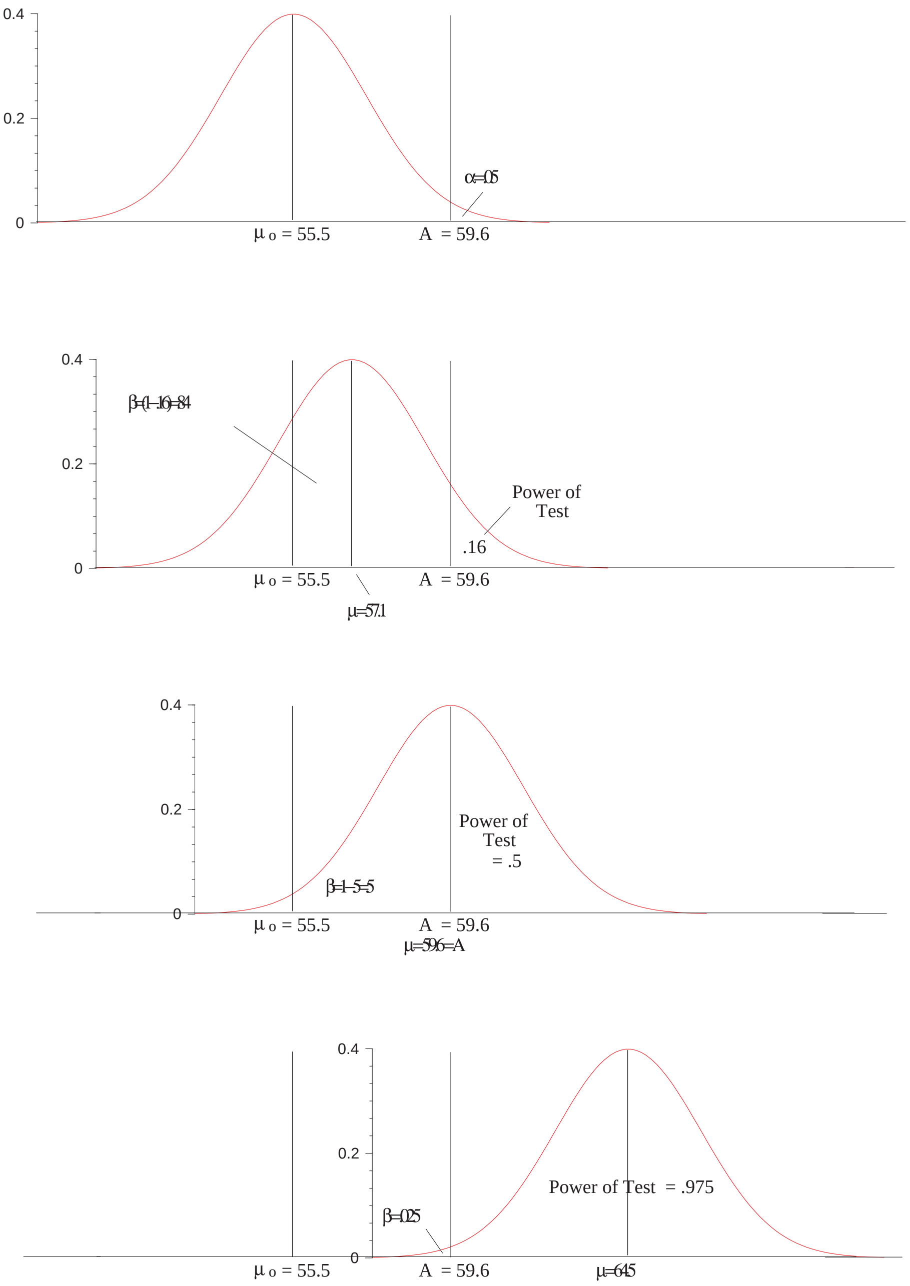

Figure 5.3: Power of test at different values of μ.

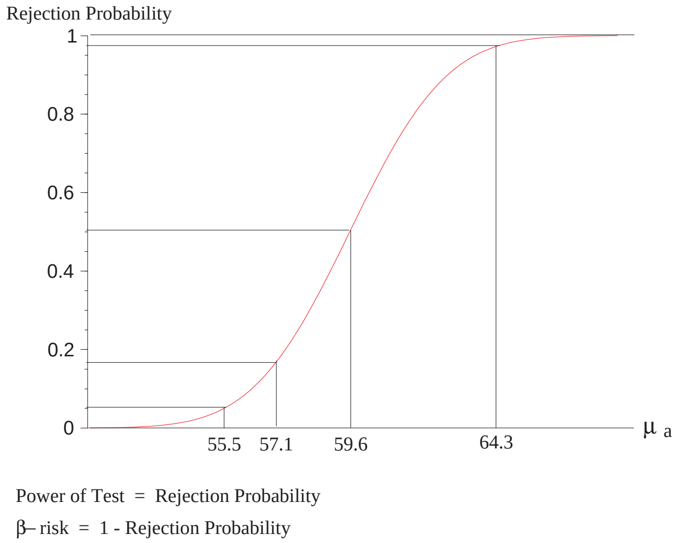

Figure 5.4: Rejection probabilities, β-risk and power of test.

that figure is called the *power curve*. The distance of that curve from the horizontal axis gives for each true value of μ the rejection probability or power of test. And the distance of the curve at each value of μ from the horizontal line at the top of the figure associated with a rejection probability of unity gives the β-risk.

The problem is, of course, that we do not know the actual value of μ (if we did, the test would be unnecessary). We thus have to choose the value of μ that we want to use to control for the β-risk. If we choose $\mu = 64.5$ as that value we can say that the power of test is .975 at μ equal to 64.5.

It can easily be seen from Figure 5.3 that the higher the value we set for the α-risk, the lower will be the β-risk at every value of μ we could set to control for the β-risk. A higher level of α will result in a critical value A closer to μ_0. The further to the left is the vertical line A, the bigger will be the power of test and the smaller will be the β-risk at every control value for μ.

The above illustration of the power of test is for one-sided upper-tail tests. For one-sided lower-tail tests the analysis is essentially the same except that A is now on the opposite side of μ_0. To portray the results graphically,

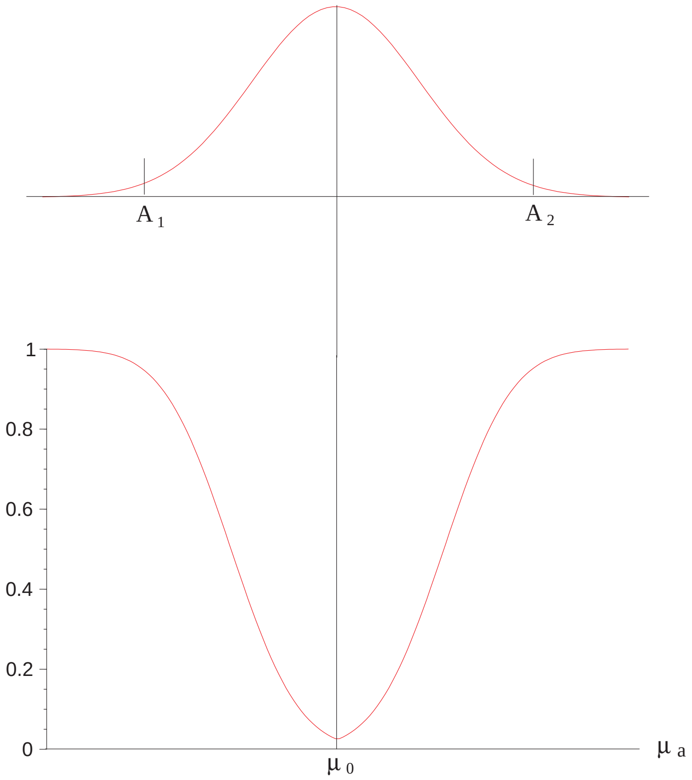

Figure 5.5: Two-sided Rejection probabilities.

simply use the mirror images of the panels in figures 5.3 and 5.4. In the case of two-sided tests the situation is a bit more complicated. The power curve now has two branches as can be seen in Figure 5.5. For any given level of μ selected to control for the β-risk the power of test will be the sum of the areas of the distribution of the sample statistic, now centered around that value of μ, to the left and right of the fixed critical levels A_1 and A_2 respectively. As the control value of μ deviates significantly from μ_0 in either direction, however, only the tail of the distribution in that direction remains relevant because the critical area on the other tail becomes miniscule. The power of test for hypotheses about population proportions is determined in exactly the same manner as above, except that the controls for the α-risk and β-risk are values of p instead of values of μ.

5.7 Planning the Sample Size to Control Both the α and β Risks

We have shown above that the lower the α-risk, the higher will be the β-risk at every level of μ at which the β-risk can be controlled. And the higher that control value of μ the greater will be the power of test.

To simultaneously control both the α-risk and the β-risk we have to choose an appropriate sample size. To choose the appropriate sample size (which must in any case be reasonably large and a small fraction of the size of the population) we must specify three things. We must specify the value μ_0 at which the α-risk is to be controlled, the value of μ, call it μ_a, at which the β-risk is to be controlled, and the planning value of σ.

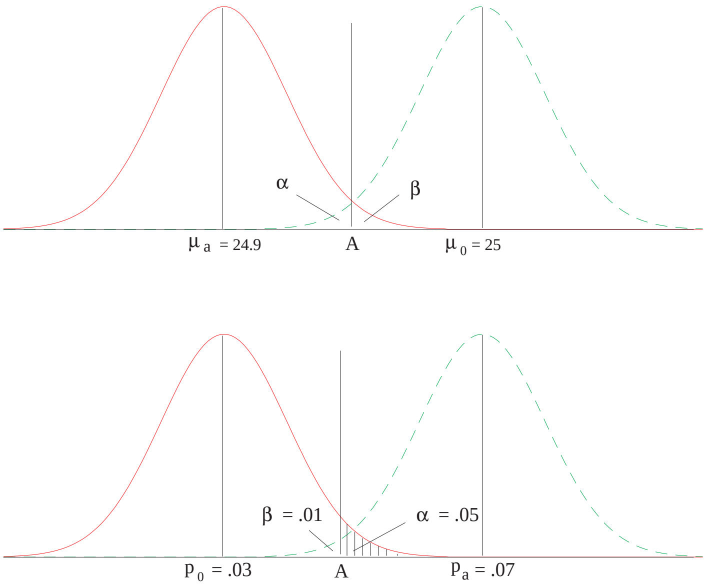

Figure 5.6: Selection of optimal sample size in butter purchase agreement problem (top) and tile shipment problem (bottom).

As a practical example, consider a purchase agreement between an aid agency and a group of producers of agricultural products. The agreement stipulates a price per 25-kilogram box of butter, but it is in the shipping company's interest to make the boxes light. A sample is to be taken from each shipment by an independent inspection group to test whether the mean weight of butter per box is 25 kilograms. The seller does not want shipments rejected if the mean weight of butter per box is equal to or above 25 kilograms. The agreement thus specifies that null-hypothesis be H_0: $\mu \geq 25$, making the alternative hypothesis H_1: $\mu < 25$. The α-risk of rejecting a shipment when mean weight of the shipment is at least 25 kilograms is set at .05. At the same time, the company purchasing the butter is interested in the boxes not being too underweight. So the agreement also stipulates that there be no more than a five percent chance that a shipment will be accepted if it contains less than 24.9 kilograms per box of butter. This controls the β-risk of erroneously rejecting the alternative hypothesis at .05 for a value of $\mu = \mu_a = 24.9$. The problem then is to choose a sample size such that the examination process will control the α-risk at .05 when $\mu = 25$ and the β-risk at .05 when $\mu = 24.9$. The buyer and seller agree to adopt a planning value of σ equal to .2. The analysis can be illustrated with reference to the upper panel of Figure 5.6. Let the as yet to be determined critical value for rejection of a shipment be A. The standardised difference between μ_0 and A must equal

$$z_0 = \frac{\mu_0 - A}{\sigma/\sqrt{n}} = \frac{25 - A}{.2/\sqrt{n}} = 1.645$$

and the standardised difference between A and μ_a must equal

$$z_1 = \frac{A - \mu_a}{\sigma/\sqrt{n}} = \frac{A - 24.9}{.2/\sqrt{n}} = 1.645.$$

These expressions can be rewritten as

$$25 - A = (1.645)(.2/\sqrt{n})$$

and

$$A - 24.9 = (1.645)(.2/\sqrt{n}).$$

Adding them together yields

$$25 - 24.9 = .1 = (2)(1.645)(.2/\sqrt{n}) = (3.29)(.2)/\sqrt{n})$$

which implies that

$$n = (\sqrt{n})^2 = \left(\frac{(.2)(3.29)}{.1}\right)^2 = 43.3.$$

A sample size of 44 will do the trick. The critical value A will equal

$$25 - 1.645\frac{.2}{\sqrt{44}} = 25 - (1.645)(.0301) = 25 - .05 = 24.95.$$

Consider another example where a purchaser of a large shipment of tiles wishes to control the β-risk of accepting the shipment at .01 when the proportion of tiles that are damaged is $p = .07$ while the vendor wishes to control the α-risk of having the shipment rejected at .025 when the proportion of damaged tiles is .03. A random sample of tiles will be selected from the shipment by the purchaser on the basis of which a decision will be made to accept (H_0: $p \leq .03$) or reject (H_1: $p > .03$) the shipment. We need to find the sample size sufficient to meet the requirements of both the purchaser and vendor. The analysis can be conducted with reference to the bottom panel of Figure 5.6. The standardised distance between the as yet to be determined critical value A and $p = .03$ must be

$$z_0 = \frac{A - .03}{\sigma_{\bar{p}_0}} = \frac{A - .03}{\sqrt{(.03)(.97)/n}} = 1.96$$

and the standardised difference between .07 and A must be

$$z_1 = \frac{.07 - A}{\sigma_{\bar{p}_1}} = \frac{.07 - A}{\sqrt{(.07)(.93)/n}} = 2.326.$$

Note that we use the values of p at which the α-risk and β-risk are being controlled to obtain the relevant values of $\sigma_{\bar{p}}$ for standardizing their differences from the critical value A. Multiplying both of the above equations by $\sqrt{n}$ and then adding them, we obtain

$$(.04)\sqrt{n} = (1.96)\sqrt{(.03)(.97)} + (2.326)\sqrt{(.07)(.93)}$$

which yields

$$n = \left(\frac{(1.96)\sqrt{(.03)(.97)} + (2.326)\sqrt{(.07)(.93)}}{.04}\right)^2 = 538.$$

The critical value A will then equal

$$A = .03 + z_0\,\sigma_{\bar{p}_0} = .03 + 1.96\,\sqrt{(.03)(.97)/538} = .0444.$$

5.8 Exercises

1. It is desired to test H_0: $\mu \geq 50$ against H_1: $\mu < 50$ with a significance level $\alpha = .05$. The population in question is normally distributed with known standard deviation $\sigma = 12$. A random sample of $n = 16$ is drawn from the population.

a) Describe the sampling distribution of $\bar{X}$, given that $\mu = 50$.

b) If μ is actually equal to 47, what is the probability that the hypothesis test will lead to a Type II error. (.74059)

c) What is the power of this test for detecting the alternative hypothesis H_a: $\mu = 44$? (.5213)

2. A sales analyst in a firm producing auto parts laboriously determined, from a study of all sales invoices for the previous fiscal year, that the mean profit contribution per invoice was \$16.50. For the current fiscal year, the analyst selected a random sample of 25 sales invoices to test whether the mean profit contribution this year had changed from \$16.50 (H_1) or not (H_0). The sample of 25 invoices yielded the following results for the invoice profit contributions: $\bar{X} = \$17.14$, $s = \$18.80$. The α risk is to be controlled at 0.05 when $\mu = 16.50$.

a) Conduct the test. State the alternatives, the decision rule, the value of the standardised test statistic, and the conclusion.

b) What constitute Type I and Type II errors here? Given the conclusion above, is it possible that a Type I error has been made in this test? Is a Type II error possible here? Explain.

3. In a tasting session, a random sample of 100 subjects from a target consumer population tasted a food item, and each subject individually gave it a rating from 1 (very poor) to 10 (very good). It is desired to test H_0: $\mu \leq 6.0$ vs. H_1: $\mu > 6.0$, where μ denotes the mean rating for the food item in the target population. A computer analysis of the sample results showed that the one-sided P–value of the test is .0068.

a) Does the sample mean lie above or below $\mu_0 = 6.0$?

b) What must be the value of value of z generated by the sample? (2.47)

c) The sample standard deviation is $s = 2.16$. What must be the sample mean $\bar{X}$? (6.5332)

d) Does the magnitude of the P–value indicate that the sample results are inconsistent with conclusion H_0? Explain.

4. The developer of a decision-support software package wishes to test whether users consider a colour graphics enhancement to be beneficial, on balance, given its list price of $800. A random sample of 100 users of the package will be invited to try out the enhancement and rate it on a scale ranging from -5 (completely useless) to 5 (very beneficial). The test alternatives are H_0: $\mu \leq 0$, H_1: $\mu > 0$, where μ denotes the mean rating of users. The α risk of the test is to be controlled at 0.01 when $\mu = 0$. The standard deviation of users' ratings is $\sigma = 1.3$.

a) Show the decision rule for $\bar{X}$ relevant for this test.

b) Calculate the rejection probabilities at $\mu = 0, 0.5, 1.0$ and 1.5 for the decision rule above.

c) Sketch the rejection probability curve for the decision rule you selected above.

d) What is the incorrect conclusion when $\mu = 0.60$? What is the probability that the above decision rule will lead to the incorrect conclusion when $\mu = .60$? Is the probability an α or β risk?

5. "Take the Pepsi Challenge" was a marketing campaign used by the Pepsi-Cola Company. Coca Cola drinkers participated in a blind taste test where they were asked to taste unmarked cups of Pepsi and Coke and were asked to select their favourite. In one Pepsi television commercial the announcer states that "in recent blind taste tests more than half the Diet Coke drinkers surveyed said they preferred the taste of Pepsi." (*Consumer's Research*, May 1993). Suppose that 100 Coke drinkers took the Pepsi challenge and 56 preferred the taste of Diet Pepsi. Does this indicate that more than half of all Coke drinkers prefer the taste of Pepsi?

6. A salary survey conducted on behalf of the Institute of Management Accountants and the publication *Management Accounting* revealed that the average salary for all members of the Institute was $56,391. A random

sample of 122 members from New York were questioned and found to have a mean salary of \$62,770 and a standard deviation of $s = \$28972$ (*Management Accounting*, June 1995).

a) Assume that the national mean is known with certainty. Do the sample data provide sufficient evidence to conclude that the true mean salary of Institute members in New York is higher than the National Average?

b) Suppose the true mean salary for all New York members is \$66,391. What is the power of your test above to detect this \$10,000 difference?

7. One of the most pressing problems in high-technology industries is computer-security. Computer security is typically achieved by a *password*—a collection of symbols (usually letters and numbers) that must be supplied by the user before the computer system permits access to the account. The problem is that persistent hackers can create programs that enter millions of combinations of symbols into a target system until the correct password is found. The newest systems solve this problem by requiring authorized users to identify themselves by unique body characteristics. For example, system developed by Palmguard, Inc. tests the hypothesis

H_0: The proposed user is authorized

versus

H_1: The proposed user is unauthorized.

by checking characteristics of the proposed user's palm against those stored in the authorized users' data bank (*Omni, 1984*).

a) Define a Type I error and a Type II error for this test. Which is the more serious error? Why?

b) Palmguard reports that the Type I error rate for its system is less than 1% where as the Type II error rate is .00025%. Interpret these error rates.

c) Another successful security system, the EyeDentifyer, "spots authorized computer users by reading the one-of-a-kind patterns formed by the network of minute blood vessels across the retina at the back of the eye." The EyeDentifier reports Type I and Type II error rates of .01% (1 in 10,000) and .005% (5 in 100,000), respectively. Interpret these rates.

8. Under what circumstances should one use the t-distribution in testing an hypothesis about a population mean? For each of the following rejection regions, sketch the sampling distribution of t, and indicate the location of the rejection region on your sketch:

a) $t > 1.440$ where $v = 6$.

b) $t < -1.782$ where $v = 12$.

c) $t < -2.060$ or $t > 2.060$ where $v = 25$.

9. Periodic assessment of stress in paved highways is important to maintaining safe roads. The Mississippi Department of Transportation recently collected data on number of cracks (called *crack intensity*) in an undivided two-lane highway using van-mounted state-of-the-art video technology (*Journal of Infrastructure Systems*, March 1995). The mean number of cracks found in a sample of eight 5-meter sections of the highway was $\bar{X} = .210$, with a variance of $s^2 = .011$. Suppose that the American Association of State Highway and Transportation Officials (AASHTO) recommends a maximum mean crack intensity of .100 for safety purposes. Test the hypothesis that the true mean crack intensity of the Mississippi highway exceeds the AASHTO recommended maximum. Use $\alpha = .01$.

10. Organochlorine pesticides (OCP's) and polychlorinated biphenyls, the familiar PCB's, are highly toxic organic compounds that are often found in fish. By law, the levels of OCP's and PCB's in fish are constantly monitored, so it is important to be able to accurately measure the amounts of these compounds in fish specimens. A new technique called matrix solid-phase dispersion (MSPD) has been developed for chemically extracting trace organic compounds from solids (*Chromatographia*, March 1995). The MSPD method was tested as follows. Uncontaminated fish fillets were injected with a known amount of OCP or PCB. The MSPD method was then used to extract the contaminant and the percentage of the toxic compound uncovered was measured. The recovery percentages for $n = 5$ fish fillets injected with the OCP Aldrin are listed below:

99 102 94 99 95

Do the data provide sufficient evidence to indicate that the mean recovery percentage of Aldrin exceeds 85% using the new MSPD method? Set the α-risk at .05.